

Lesson 3.2.1

Using Lines to Make Predictions



INSTRUCTOR SPECIFIC MATERIAL IS INDENTED AND APPEARS IN GREY

ESTIMATED TIME

50 minutes

MATERIALS REQUIRED

Overhead or electronic display of scatterplots in lesson

BRIEF DESCRIPTION

This lesson lays the foundation for understanding bivariate relationships and using scatterplots with summary lines to make predictions.

LEARNING GOALS**Students will understand that:**

- In a statistical relationship two variables tend to vary together in a predictable way. When a line is a good summary of a statistical relationship, the line can be used to predict the value of a response variable given a value of the explanatory variable (the predictor), but only part of the variation in y is explained by changes in x .
- Predictions are more accurate when the relationship is strong.
- Making predictions based on extrapolating outside the range of the data can be risky and should be done with caution.

Students will be able to:

- Identify the response variable (dependent variable) and the explanatory variable (predictor variable) given a scenario involving a bivariate numerical data set

Lesson 3.2.1

Using Lines to Make Predictions

- Predict the value of the response variable using both the graph of the line and its equation given a regression line and a value for the predictor variable
- Explain the danger of extrapolation in a regression setting

INTRODUCTION

The focus of this lesson is on whether a line seems to be a reasonable summary of the bivariate relationship and, if it is, using a line to predict y based on x . Students will probably begin to sketch lines onto the scatterplots to make their predictions before this idea is formally introduced in the lesson. In this lesson, students use both the graph and the equation of the least squares regression line to make predictions. In subsequent lessons in this topic, you will review the concepts of slope and y -intercept explicitly. You will also delve into the idea of “best fit” based on the sum of the squares of residuals, so it is not necessary to teach these concepts here. However, take opportunities if they arise from students’ comments to connect to students’ knowledge of linear functions.



STUDENT MATERIAL IS NOT INDENTED AND APPEARS IN BLACK

INTRODUCTION

Statistical methods are used in forensics to identify human remains based on the measurements of bones. In the 1950s, Dr. Mildred Trotter and Dr. Goldine Gleser measured skeletons of people who had died in the early 1900s. From these measurements they developed statistical methods for predicting a person’s height based on the lengths of various bones. These formulas were first used to identify the remains of U.S. soldiers who died in World War II and were buried in unmarked graves in the Pacific zone. Modern forensic scientists have made adjustments to the formulas developed by Trotter and Gleser to account the differences in bone length and body proportions of people living now. You will not use Trotter and Gleser’s formulas in this problem, but you will use a similar process.

Note: For information on the Terry skeleton collection, see <http://anthropology.si.edu/cm/terry.htm>. For a more recent example of how forensic scientists are still building on the work of Trotter and Gleser, see the following:

- Jantz, R. L. (1993). Modification of the Trotter and Gleser female stature estimation formulae. *Journal of Forensic Science*. 38(4), 758–63.)

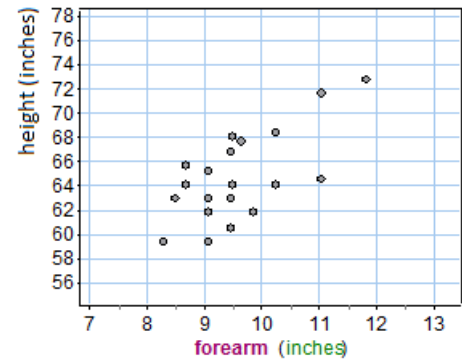
Lesson 3.2.1

Using Lines to Make Predictions

To illustrate the type of data analysis done in forensics, let's see if you can identify a female student based on the length of her forearm. The mystery student has a forearm measurement of 10 inches. (She is alive and healthy!)

Height and weight measurements for three female college students are given in the table.

	Jane Doe 1	Jane Doe 2	Jane Doe 3
Age	18	23	33
Gender	Female	Female	Female
Height	5 feet, 5 inches	5 feet, 2 inches	6 feet
Weight	128 pounds	120 pounds	155 pounds



Your task is to determine if the mystery student could be one of these three students.

First, you need data that relates forearm length to either height or weight for females. The scatterplot is a graph of height versus forearm length for 21 female college students taking Introductory Statistics at Los Medanos College in Pittsburg, California, in 2009.

Note: Let students work for a few minutes alone, and then compare responses with a neighbor or group. Students may feel uncomfortable making a prediction since there is clearly no “right” answer. This is acceptable. Remind them that in statistics they are constantly making decisions in the face of variability in the data. Perhaps also give an obviously unreasonable prediction, such as a height of 56 or 76, ask students if the prediction seems reasonable, and then encourage them to give a better prediction.

- 1 Based on the scatterplot, what is a reasonable prediction for the height of the mystery student? Briefly explain or show how you made your prediction.

Answer: Will vary.

- 2 The variability in the data makes it difficult to determine if one of these students is the mystery student. Could any of the three students be eliminated as a possibility of being the mystery student? Explain your reasoning.

Answer: Will vary but students should be able to eliminate one.

Lesson 3.2.1

Using Lines to Make Predictions

WRAP-UP

Based on the data in the scatterplot, ask students to determine if the following predictions for the height of the mystery student are reasonable or unreasonable (perhaps ask students to show a thumbs up for *reasonable prediction* and thumbs down for *unreasonable prediction*): 62 inches (reasonable), 65 inches (reasonable), 74 inches (unreasonable).

Plot each of these predictions on the scatterplot and highlight how the prediction fits the pattern in the data or deviates from the pattern. Call on a few students to give their predictions and plot these as well.

Indicate with a vertical line segment a reasonable range of predictions for $x = 10$ (approximately 62–68 inches). Based on this range of reasonable predictions, it looks like the mystery student is probably not Jane Doe 3, who is 72 inches tall.

Most likely students have already started eyeballing lines to fit the data, so you may want to elicit this idea before moving into Part II. You could do this by asking a few students to explain how they determined their prediction or just point out that the association is positive and somewhat linear, and then ask if anyone used a line to help summarize the data to make a prediction. If so, congratulate them for thinking like a statistician!

INTRODUCTION

In **Next Steps**, students are introduced to making predictions using both the equation and graph of a line. The concepts in this lesson will probably not be difficult for most students, so the wrap-up for this segment of the lesson is an assessment item.

Decide whether you will use the problems in **Next Steps** as group work, to frame a whole-class discussion, or as a basis for a lecture. Orient students accordingly. Since the wrap-up does not include a summary, if you are conducting the lesson using group work, address individual difficulties as students are working.

This should take students 20–30 minutes total, including about 6 minutes for assessment item at the end of the lesson

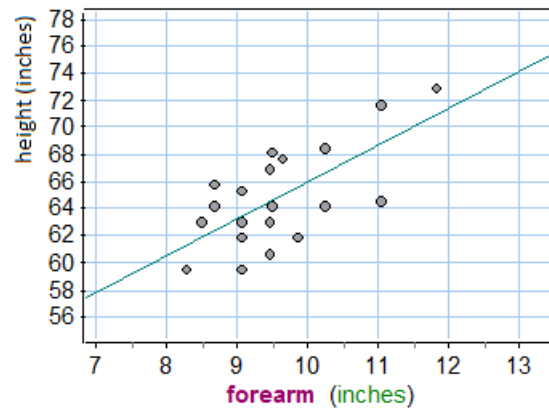
NEXT STEPS

Lesson 3.2.1

Using Lines to Make Predictions

Using a Line to Make Predictions

- 3 The scatterplot has a positive linear association. The correlation is 0.68, which is pretty strong. So, it makes sense to use a linear model to summarize the relationship between the forearm and height measurements. There is one line that is considered the best description of how height and forearm length are related. You will learn more about how to find this line in future lessons. For now, you will use technology to find the equation of this line.



- A Use the graph of the best-fit line to predict the height of the mystery student.

Answer: 66 inches.

- B The equation of this line is approximately predicted height = $2.7(\text{forearm length}) + 39$.

$$\hat{y} = 2.7x + 39$$

(Notice that when you use letters to represent variables in the prediction line, you put a “hat” on the y and write \hat{y} instead of y . The hat is a signal that the variable is *predicted* values, not actual data values.)

Use the equation to predict the height of the mystery person.

Answer: 66 inches.

- C Is the height of Jane Doe 1, 2, or 3 closest to the predicted height of the mystery student given by the line? (Of course, this does not guarantee that you have correctly identified the mystery student, but it suggests that one student’s height, together with the 10-inch forearm measurement, fits the linear pattern in the data better than the other students.)

Answer: Jane Doe 1.

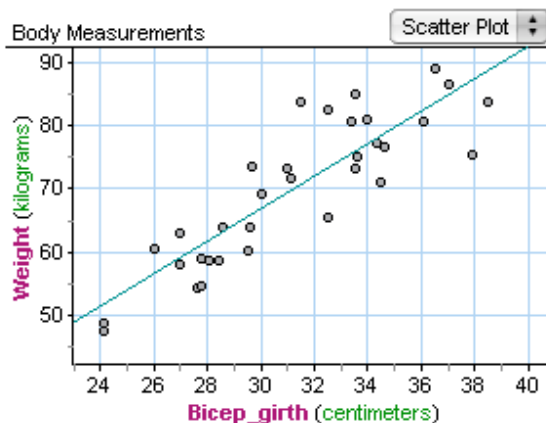
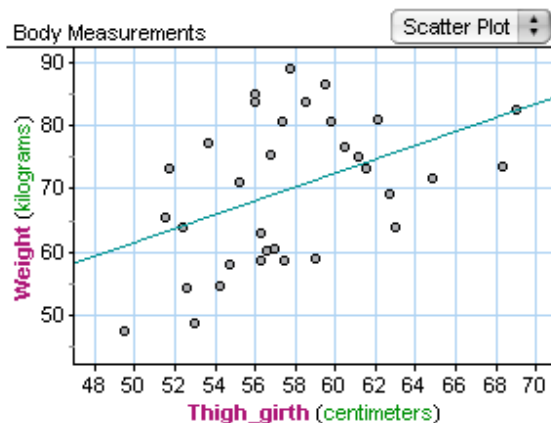
- 4 The scatterplots below are graphs of body measurements in centimeters for 34 adults who are physically active. These data are a random sample taken from a larger nonrandom data set gathered by researchers investigating the relationship of various body measurements and weight.

Lesson 3.2.1

Using Lines to Make Predictions

Girth is the measurement around a body part.

(Retrieved from www.amstat.org/publications/jse/v11n2/datasets.heinz.html)



- A Based on these data, which do you think is a better predictor of an adult's weight: thigh girth or bicep girth? Why?

Answer: Bicep girth because there is less scatter about the line.

- B Adriana has a thigh girth of 57 centimeters and a bicep girth of 25 centimeters. Predict her weight using the measurement that you think will give the most accurate prediction, and then plot Adriana on the scatterplot that you used to make her weight prediction.

Answer: Use the bicep girth. 54 kilograms is a reasonable prediction from line.

- C The equations of the two lines shown are

$$\text{weight} = 6.3 + 1.1(\text{thigh girth}) \quad \text{weight} = -10.5 + 2.6(\text{bicep girth})$$

Predict Adriana's weight using the equation that you think best predicts weight.

Answer: 54.5 kilograms.

- D Of course, you do not really know Adriana's weight. How accurate do you think the line's prediction of Adriana's weight is? Choose the option that is the most reasonable and explain your thinking.

- Very accurate (within a range of plus or minus 1 kilogram).
- Somewhat accurate (within a range of plus or minus 5 kilograms).

Lesson 3.2.1

Using Lines to Make Predictions

- Not very accurate (within a range of plus or minus 10 kilograms).

Answer: Somewhat accurate. One way to see this is to shade a region parallel to the line with width ± 5 kilograms to show that most of the data falls within this range of the predicted values.

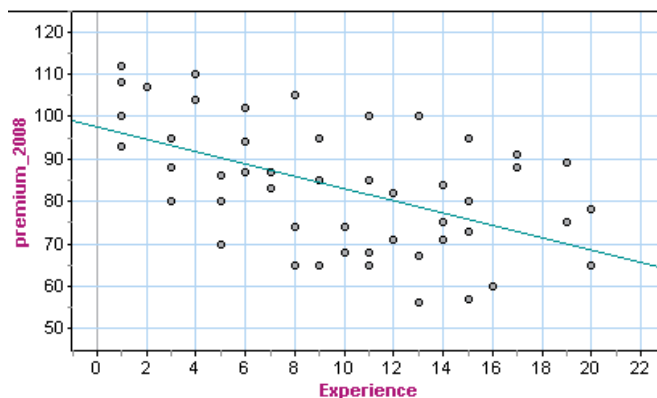
- 5 In previous lessons, you studied the concept of correlation to describe the strength and direction of the linear association between two quantitative variables. Now you are working on predicting the value of one variable based on the other. Are these two ideas related? Explain your reasoning.

Answer: Yes. If the relationship looks linear, a strong association, indicated by r close to 1 or -1 , indicates that there is not a lot of scatter about the line. So predictions will probably be more accurate.

TRY THESE

In 2008, a statistics student gathered data on monthly car insurance premiums paid by students and faculty at Los Medanos College. Relating monthly car insurance premiums to years of driving experience, she found a linear relationship and used statistical methods to get the following equation:

$$\text{predicted monthly car insurance premium} = 97 - 1.45(\text{years of driving experience})$$



- 6 Predict the monthly car insurance premium paid by someone who has been driving 12 years.

Answer: \$80.

- 7 Which of the following methods can be used to make the prediction?

Lesson 3.2.1

Using Lines to Make Predictions

- A Find 12 on the horizontal axis, trace up to the line, and read off the corresponding value on the y -axis.
- B Substitute 12 in the equation, and calculate the predicted premium.
- C Look at the data and find a person who has been driving 12 years. Report the premium paid by this person.
- D Both A and B.
- E Both B and C.

Answer: D.

WRAP-UP

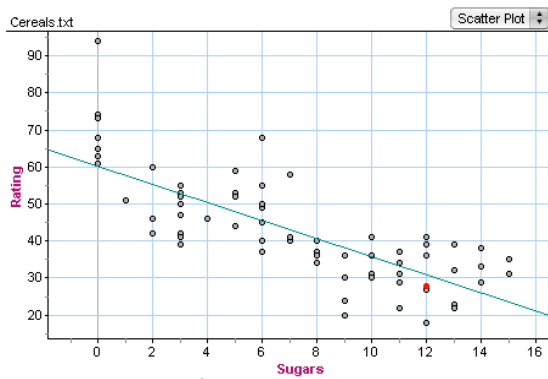
Give students about 3-5 minutes to do the **Try These** exercises. Assess class performance in aggregate protecting anonymity by using clickers or have students write their answer anonymously on a piece of paper, collect, redistribute, and tally responses with students reporting the answer on the paper they receive. Class results will guide you in determining whether more explanation is necessary.

Lesson 3.2.1

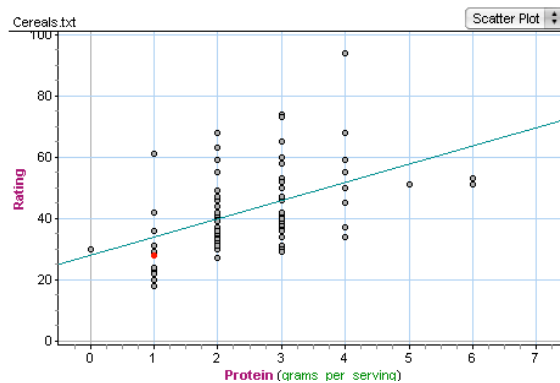
Using Lines to Make Predictions

TAKE IT HOME

- Here you return to the data set for the 77 breakfast cereals you investigated at the beginning of Module 3.



$$\text{ratings} = 60 - 2.43(\text{sugars})$$



$$\text{ratings} = 8 + 5.96(\text{protein})$$

Two new cereals are being rated by *Consumer Reports*. Cereal A has 10.5 grams of sugar in a serving and Cereal B has 2.5 grams of protein in a serving.

- A Predict the *Consumer Reports* rating for the two cereals using the best-fit lines.

Answer: Cereal A: 34.485, Cereal B: 22.9. Note that the line does not give whole number ratings. Students may round to indicate a realistic rating value.

- B For which cereal do you think your prediction is probably more accurate (more likely to be closer to the actual *Consumer Reports* rating)? Why?

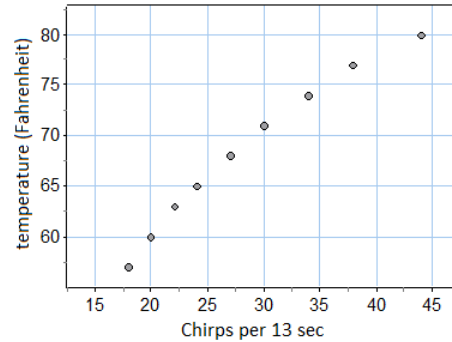
Answer: The prediction for Cereal A is probably more accurate because there is less scatter about the line.

Lesson 3.2.1

Using Lines to Make Predictions

- 2 Can the rate that crickets chirp be used to predict the temperature?

According to Tom Walker, an entomologist with the University of Florida, all crickets are pretty good thermometers because they chirp at a rate that is related to the temperature. The chirping noise results when the cricket rubs its wings together. A cricket studied by Walker, the snowy tree cricket (*Oecanthus fultoni*), chirps at a rate that is slow enough to count. These crickets also synchronize their wing rubbing so determining the chirp rate easier. The snowy tree cricket is found throughout the United States. To hear the snowy tree cricket go to <http://entnemdept.ufl.edu/walker/buzz/585a.htm>.



- A The scatterplot is a graph of data from the June 1995 issue of *Outside* magazine. Use the scatterplot to predict the temperature when the snowy tree crickets are chirping at a rate of 40 chirps every 13 seconds.

Answer: $77^{\circ}\text{F} \pm 1^{\circ}\text{F}$ is a reasonable estimate.

- B How accurate do you think your prediction is? Choose the option that is most reasonable and briefly explain your thinking.

- Very accurate (within a range of plus or minus 1 degree).
- Somewhat accurate (within a range of plus or minus 5 degrees).
- Not very accurate (within a range of plus or minus 10 degrees).

Answer: Very accurate. The association is so strong you can sketch a curve that has essentially no scatter about it.

- C This is the same data graphed in two different windows. The data has been zoomed out by expanding both axes. The line pictured is the best-fit line:

$$\text{temperature} = 0.88(\text{chirp rate}) + 43$$

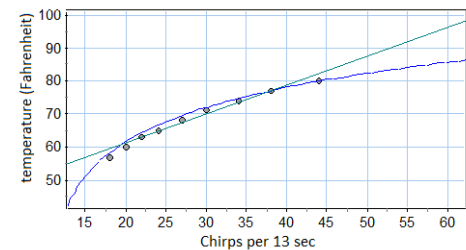
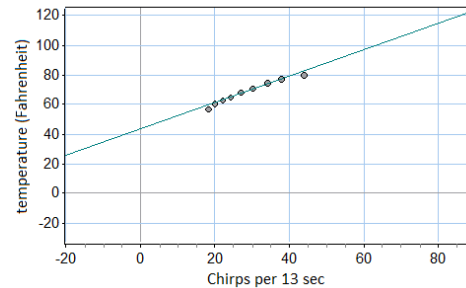
For some chirp rates, this line gives very accurate predictions of the temperature. However, the data are actually slightly curved, so that for chirp rates above 50 a nonlinear model might give more accurate predictions. One possible nonlinear model is also shown.

Lesson 3.2.1

Using Lines to Make Predictions

The line also has limitations in that some chirp rates are meaningless and should not be used to make predictions.

In statistics, *extrapolation* is the process of using a statistical model (like a line) to make predictions outside the range of the available data. To use a statistical model to make a prediction for an explanatory variable value that is outside the range of values in the data set requires that we make the assumption that the pattern observed in the data continues outside this range. If this is not the case, predictions are unreliable and may be very far off from the actual response variable values. You should be very cautious in doing this.



Illustrate the concept of extrapolation by identifying a point on the line that gives either meaningless results or unreliable results. Explain how this point illustrates the concept of extrapolation.

Answers can vary: Answers need to use chirp rates outside the range of the data to illustrate extrapolation. For example, using the linear model, a chirp rate of 60 gives a prediction of 95.8°F. However, the curve in the data suggests that a prediction of 90 might be more accurate. Another example is using a chirp rate of -10, which is a nonsensical value, and predicting a temperature of 30°F, a temperature at which crickets are probably dead.

Lesson 3.2.1

Using Lines to Make Predictions

- 3 **A Note About Statistical Vocabulary:** A variable that is used to predict the value of another variable is called the **predictor variable**, also known as the **independent variable** or **explanatory variable**. The other variable, whose values you are predicting, is called the **response variable**, also known as the **dependent variable**.

- A The introductory problem in this lesson has forearm lengths and heights for 21 female college students. In this situation, which variable is the predictor?

| **Answer:** Forearm lengths.

- B The cereal data has the amount of sugar in a serving and the *Consumer Reports* rating. In this situation, which variable is the predictor?

| **Answer:** Sugar.

- C When graphing bivariate data, you put the predictor variable on the (*choose one*: horizontal axis, vertical axis).]

| **Answer:** Horizontal axis.

- D Using measurements of temperature (°F) and the chirp rate of the snowy tree cricket (measured in number of chirps in 13 seconds), students use technology to find a best-fit line. However, some students use temperature as the predictor variable, and others use chirp rate as the predictor variable. For which of the two lines below is temperature treated as the predictor variable?

$$\text{temperature} = 0.88(\text{chirp rate}) + 43$$

$$\text{chirp rate} = 1.1(\text{temperature}) - 47$$

| **Answer:** Chirp rate = $1.1(\text{temperature}) - 47$

+++++

This lesson is part of STATWAY™, A Pathway Through College Statistics, which is a product of a Carnegie Networked Improvement Community that seeks to advance student success. Version 1.0, A Pathway Through Statistics, Statway™ was created by the Charles A. Dana Center at the University of Texas at Austin under sponsorship of the Carnegie Foundation for the Advancement of Teaching. This version 1.5 and all subsequent versions, result from the continuous improvement efforts of the Carnegie Networked Improvement Community. The network brings together community college faculty and staff, designers, researchers and developers. It is an open-resource research and development community that seeks

Lesson 3.2.1

Using Lines to Make Predictions

to harvest the wisdom of its diverse participants in systematic and disciplined inquiries to improve developmental mathematics instruction. For more information on the Statway Networked Improvement Community, please visit **carnegiefoundation.org**. For the most recent version of instructional materials, visit **Statway.org/kernel**.

+++++

STATWAY™ and the Carnegie Foundation logo are trademarks of the Carnegie Foundation for the Advancement of Teaching. A Pathway Through College Statistics may be used as provided in the CC BY license, but neither the Statway trademark nor the Carnegie Foundation logo may be used without the prior written consent of the Carnegie Foundation.